Weekly Report

2013-6-9

Feiran Wu

1 Introduction

The idea combines time series data with text analysis and visualization has been listed this week. I have investigated the relevant algorithms and methods then made an initial plan.

2 RESEARCH

The First step is to transfer time series data to alphabetic string. The parameters in this process are about the limitation of alphabet. Which means we need a reference that could define how many levels should be set when discretize time series. The main reference about this step is SAX.

Next step need to segment parts of string to extract words in each time series. There are two strategies to deal with different situations:

1. When dealing the single time series independently, we could first collect the N-grams substrings that arise with highest frequency. This step may extract the ordinary or periodic words. We may use these words to construct a model that reflects the normal situation of this time series. According this model we may pick up some motif words when import more data (time series in future). To summarize, we could extract ordinary (appear frequently), periodic (appear in a certain period) and motif (abnormal) words.
2. When dealing the multi-dimensional time series. We suppose some of them are dependent. We may use the method listed in the previous (single time series) to extract words of single dimension then comparing them and get the common words. Or we can cluster time series in different dimension into different group (e.g. in Gulf of Maine Ocean data, the ocean currents magnitude in different depth may be clustered together, water temperature may be clustered together. Specially, salinity may be clustered into water temperature group since they indeed show some positive correlation relationship although I don't know the reason). Then use multiple sequences alignment method to extract some common features (words). The purpose of words extraction of multi-dimensional time series is to get the words in one group. In other words, we can gain some words describe the nature of one group.

The third step is to merge words into event and explore the event relationship. For example, if event A always happen before event B, we may establish causality between them. If event A always happen with event B simultaneously, we may establish commensalism between them.

The last step is to show these relationships among the events to describe the nature of dataset. The data changes pipeline has been show in Figure 1.

Figure 1 The data changes pipeline.

3 CONCLUSION

Next week’s job :

More concrete plan.

The initial implements of step 1 & 2.